REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 23012012	**2. REPORT TYPE** Final Technical	**3. DATES COVERED** *(From - To)* 03012007-31082011

4. TITLE AND SUBTITLE Speech Synthesis Using Perceptually Motivated Features	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER FA9550-07-1-0199
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Dr. Steven Greenberg	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Silicon Speech Speech Technology Research 46 Oxford Dr San Rafael, CA 94903-2886	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR 875 North Randolph St Arlington, VA 22203	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-OSR-VA-TR-2012-0151

12. DISTRIBUTION/AVAILABILITY STATEMENT

a - Approved For Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The trajectory of this project parallels, in certain ways, the changing dynamics of speech and neuroscience research. When the project began in 2007, many aspects of these fields were formulated in concepts and methods originating more than 50 years ago. By the project's conclusion, in May 2011, the focus in both fields had shifted to statistical (often Bayesian) approaches, with a clear recognition that the classical models require serious revision. Whether the Bayesian framework turns out to be the "best" one is uncertain. Despite its limitations (narrow perspective, lack of explanatory insight), it represents a significant improvement over traditional, quasi- deterministic approaches that dominated speech and brain research over much of the 20th century.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Steven Greenberg
					19b. TELEPHONE NUMBER *(Include area code)* 415-472-2000

20120918103

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Grant FA9550-07-1-0199

Speech Synthesis Using Perceptually Motivated Features – Final Report

Introduction

The trajectory of this project parallels, in certain ways, the changing dynamics of speech and neuroscience research. When the project began in 2007, many aspects of these fields were formulated in concepts and methods originating more than 50 years ago. By the project's conclusion, in May 2011, the focus in both fields had shifted to statistical (often Bayesian) approaches, with a clear recognition that the classical models require serious revision. Whether the Bayesian framework turns out to be the "best" one is uncertain. Despite its limitations (narrow perspective, lack of explanatory insight), it represents a significant improvement over traditional, quasi- deterministic approaches that dominated speech and brain research over much of the 20th century.

The theoretical perspective of the current project shares many properties in common with the Bayesian framework, but differs in a number of significant respects. The areas of concordance and disagreement are discussed later in this report. Common to both is the increasing recognition that models and theories developed in the 20th century (or earlier) should be re-examined in light of recent developments in theory and experiment. In their place are approaches using highly sophisticated statistical and mathematical methods to handle the complexity and variability of the real world, and which are capable of being deployed in computational systems for solving real (as apposed to toy) problems.

This project began as one devoted to developing a new form of speech synthesis, using as its representational basis a theoretical framework known as the "complex modulation spectrum" (CMS). The CMS can trace its origins to the 1930s, when Homer Dudley (1939; 1940), a Bell Labs engineer, developed a primitive speech synthesizer, the Voder (followed by a more sophisticated version, the Vocoder). Dudley's great insight was that much of the speech signal's temporal structure could be filtered out without disrupting intelligibility as long as very slow energy modulations (below 25 Hz) were preserved. This observation was later applied to speech in noisy and reverberant environments by two Dutch researchers, Tammo Houtgast and Hermann Steeneken (1985), who found that the key modulation frequency range was 3–10 Hz. At the time, the basis for this temporal "importance" region was speculative. It was noted that syllables usually fell within this range; hence, some connection to linguistic structure was evident. However, the actual link between acoustics (i.e., the "modulation spectrum"), linguistic structure and human perception remained unclear. It would require many years of additional research to begin connecting the dots. The connections form the focus of much of the discussion that follows.

Initial Project Phase – Speech Synthesis and STRAIGHT TALK

The project initially focused on speech synthesis. The original objective was to determine whether certain properties of speech's low-frequency temporal structure could be used to generate novel utterances from pre-recorded material in a way that transcended the limitations of conventional concatenative approaches. In concatenative synthesis (usually) several hours of spoken (often read) material is recorded as the waveform pool for generating novel signals. Unit selection, typically based on either phonetic segments or transitions between segments, chooses the "optimum" waveform fragment from pre-recorded material for each point in time and stitches (i.e., concatenative) them together. Although this approach works reasonably well most of the time, it can produce unintelligible speech, and moreover is difficult to adapt to different speaking styles and expressive/emotional content. Because the modulation spectrum is closely linked to intelligibility, the intent was to use it as the basic signal representation for sculpting linguistic and expressive content in the waveform that would otherwise not be possible. Two collaborators, Les Atlas of the University of Washington, and Hideki Kawahara of Wakayama University in Japan agreed to work on modifying the latter's Vocoder system (STRAIGHT) to use CMS features. The system, known as STRAIGHT TALK, was designed to provide a linguistic and signal-processing interface to Kawahara's STRAIGHT so that it could generate novel material (up until then, STRAIGHT could only morph one utterance into another, without specifying novel content.

One of Les' graduate students, Cameron Colpitts, agreed to take on the project as the focus of his electrical-engineering masters thesis. He focused on transforming simple speech material (monosyllables and simple words) into other speech signals with different linguistic properties as a first step in developing an approach capable of generating truly novel material. Many different methods for effecting the transformation from one segment to another were tried. Of particular interest was the ability to transform stop consonants from one "place of articulation" to another (e.g., [p] > [t] or [k]; [g] > [d] or [b], etc.) because of their strategic role in word identity and lexical discriminability. To our surprise, the best method (in terms of naturalness and clarity) was one in which most of the original waveform properties were preserved and only slight changes made to the signal. This result was consistent with Kawahara's experience with STRAIGHT – make only small changes to preserve quality and intelligibility – but flew in the face of "theory" in that the modulation spectrum is associated with large variation in energy over time. A principled reason for this seeming paradox would only emerge a few years later when I visited Geoff Hinton in Toronto (discussed below).

While Cameron was working on the phonetic-feature-transformation technology, Hideki, Les and I were trying to find a way for CMS features to control STRAIGHT. The main issue was that CMS is inherently a negative-valued representation using imaginary numbers (because it deals explicitly with phase), while STRAIGHT is real-valued and effectively phase-free. This turned out to be a daunting incompatibility to resolve, and despite our best efforts, was never satisfactorily implemented. Hideki had spent many years developing STRAIGHT and didn't believe it could be changed in the way required to accommodate CMS features. Les was convinced that phase is crucial for signal representations and could not find a way to accommodate STRAIGHT's mathematical basis.

Years later (September, 2011), the reasons for the incompatibility would become evident. Although Les was correct in his insistence that modulation phase be preserved in the speech representation, he (as well as I and Hideki) failed to realize that STRAIGHT does actually incorporate phase into the speech representation, but in an idiosyncratic way. STRAIGHT requires that the spectral representation, derived from formant patterns, be synchronized to the glottal (pitch) period. This form of synthesis, known as "pitch-synchronous," essentially preserves the modulation phase but with a temporal granularity tied to the glottal period (ca. 10 ms for a male voice and ca. 5 ms for a female speaker). This is because STRAIGHT represents the original signal's (acoustic-frequency) spectrum as accurately as possible. It is, by necessity, highly smoothed, and the signal is temporally quantized into glottal periods (which are synchronized to STRAIGHT's spectral representation). Because the spectrum is so accurately represented in each glottal period, the modulation phase is accurately captured in STRAIGHT; however, the precision of its representation is limited by the requirement that the spectrum be synchronized to the glottal period. In this sense, STRAIGHT is able to represent the modulation phase, but not in the same mathematical framework as Les' CMS formulation (which requires that the mathematics be well-formed and internally consistent). If it were not for Cameron Colpitts' decision to leave academia (see below), it is likely we would have been able to meld a modified version of CMS with STRAIGHT. Ironically, I have revisited this specific issue (how to control STRAIGHT using modulation-centric features) in recent weeks as part of a project helping a language-instruction company (Transparent Language) develop novel technology for assessing students' pronunciation of foreign languages. Hence, the lessons learned during the Vocoder phase of the AFOSR project is likely to help develop technology that re-synthesizes a student's speech to facilitate and enhance pronunciation training.

Despite the success of Cameron's phonetic-transformation project, he decided academia was not in his immediate future. Instead, he joined a financial-forecasting company to focus on developing mathematical models for investing. Les was unable to find another student with Cameron's skill set and interest in speech synthesis. With the concurrence of Willard Larkin, the project's program manager, the sub-contract with the University of Washington (UW) was terminated effective December 1, 2008. I then pondered whether to seek another academic site to partner with in place of UW. Because of this unanticipated disruption to the project, Willard Larkin suggested a no-cost restructuring of the grant so as to extend the project period through May, 2011.

This extension provided invaluable time to decide what steps to take next. I discussed the possibility of collaboration with three separate sites (Sheffield and Plymouth in the UK, Carnegie-Mellon in the US), but concluded that their research goals did not coincide sufficiently with the AFOSR project to ensure a successful collaboration. I concluded it would be better to devote the remainder of the project to two other research areas closely related to projects initiated with others a few years prior (with the concurrence of the project's program manager, Willard Larkin).

The Perceptual Flow of Phonetic Information and Consonant Recognition

The first of these projects was based at the Technical University of Denmark (DTU), where I had spent a total of 12 months as a Visiting Professor over the course of 5 years (2004-2009). This project focused on how phonetic information was processed in the auditory system, especially in terms of low-frequency, energy-modulation properties. Such knowledge could be used to develop a speech synthesis system based on CMS features making a direct connection between it and the original AFOSR project focus.

All of the experiments run at DTU focused on Danish-consonant recognition. My DTU colleague, Thomas Christiansen, performed all of the data collection and much of the initial data analysis. My role was to design the experiments, help in analyzing and interpreting the data, as well as writing papers and constructing Powerpoint presentations. Thomas and I probably spent a couple of hundred hours just in discussion about the study and how to analyze the data. The writing (and re-writing) of the papers consumed a significant portion of my time.

There are 11 consonants in Danish, far fewer than in English (or most other Indo-European languages). These were combined with three different vowels ([i], [a], [u]) using a corpus developed at Aalborg University and embedded in a di-syllable, –lə (talə, milə, etc.) which followed a short carrier phrase. The listener's task was to indicate which of the 11 consonants they heard in the test syllable.

The syllables were spectrally and low-pass modulated filtered. Spectrally, each signal was band-passed through a 2/3-octave filter centered at 750 Hz, 1500 Hz or 3000 Hz. These parameters were carefully chosen after extensive pilot studies in order that the percentage of consonants correctly recognized were approximately equal across the three "slit" frequencies for all listeners (ca. 40%).

In the first part of the study, the three spectral slits were presented in isolation (i.e., 750-Hz slit alone, 1500-Hz alone, etc.) and in combination with one or two of the other slits. Consonant recognition for the three slits was ca. 90% correct, a performance level designed to allow us to deduce precisely the contribution made by each slit when combined with others. Hence, there was a 50% (absolute) dynamic range in recognition between the single and three-slit conditions.

In our view, the most appropriate way to analyze these recognition data is NOT in terms of percent correct (the conventional way of doing so), but in terms of an information-theoretic metric:

$$I(S;X) = \sum_{s.x} p_{SX} \log_2 \frac{p_s p_x}{p_{sx}}$$

where *I(S;X)* refers to the mutual information (or IT) between S (stimulus) and X (response; i.e., the number of bits transmitted), p_{sx} is the probability of stimulus, *s*, co-occurring with response *x*, p_s is the probability of stimulus *s* occurring, and p_x is the probability of response *x* occurring. This formulation is an adaptation of one originally published by Miller and Nicely (1955). It provides an efficient way to measure how well each consonant is recognized and at the same time allows us to determine how well that consonant is distinguished from the others. Among other benefits, it has a built-in compensation for response bias and also takes care of any inequality in stimulus presentation probability. As discussed below, this IT metric also allows us to deduce how spectral information pertaining to phonetic classes is combined across the frequency spectrum.

Each consonant can be decomposed into one of three phonetic-feature classes: (1) Voiced/Unvoiced, (2) Manner of Articulation (stop, nasal, fricative), and (3) Place of Articulation (anterior, central, posterior). In order for a consonant to be recognized all three features need to be decoded correctly ("Decoding" is used instead of "recognition" in this context because listeners were never asked to recognize phonetic features directly. Hence, decoding is essentially an "implicit" form of recognition upon which "explicit" recognition depends). As it turns out, these features are decoded quite differently from each other, and this observation allows us to gain great insight into how consonants (and by extension speech) is processed and ultimately understood.

The distinction among phonetic features is observed in how their information is combined across the acoustic-frequency spectrum. Two of the features, Voicing and Manner of Articulation, combine nearly linearly when two slits are added together. When a third slit is added, the amount of information gain is smaller than predicted from a linear summation (i.e., there is an effective "compression" of the gain function associated with information integrated across the frequency spectrum). In contrast, the information association with Place of Articulation information combines in super-linear fashion, such that two slits have roughly four times the amount of information as one, and three slits can contain as much as 10 times the information (as a single slit). In other words, combining information associated with Place of Articulation information is highly synergistic. What is the significance of this observation?

Although all three phonetic features are important for consonant recognition, place of articulation is particularly so. This is because it is the feature that serves to distinguish among different words more than the others and is also (paradoxically) more vulnerable to background noise and other forms of acoustic interference. It is also the feature most closely associated with visual speech-reading cues. Visual information from the motion of lips, tongue and jaw are virtually the same as Place-of-Articulation cues derived from the acoustic signal. In noisy conditions, the visual speech cues are particularly important because they compensate for the degraded nature of the acoustic signal. It is our hypothesis that the Place of Articulation cues are inherently synergistic, and that this is the reason why visual speech cues combine with the degraded acoustic cues so effectively to restore speech intelligibility and phonetic recognition under many listening conditions.

There is another important finding from this paper (Christiansen and Greenberg, 2012) worthy of note in this report. First, it is possible to quantify the amount contributed by each spectral region (i.e., slit) to phonetic-feature decoding and consonant classification through a measure known as "symmetric redundancy" (Cover and Thomas, 1990):

$$\hat{I}(X;Y) = \frac{2I(X;Y)}{H(X)+H(Y)}$$

where $\hat{I}$(X;Y) is the Symmetric Redundancy (SyR) of variables *X* and *Y*, *I(X;Y)* is the mutual information shared between the responses *X* and *Y*, *H(X)* is the entropy of *X* and *H(Y)* is entropy of *Y*. The SyR can range between 0, signifying complete independence (i.e., no correlation among responses), and 1, denoting complete dependence (i.e., (X;X) = 1).

SyR is essentially an IT metric of similarity based on the listener consonant-confusion patterns. What is of interest for this project is that the SyR associated with Voicing, Manner and Place information differ dramatically. There is a relatively high degree of SyR for Voicing and Manner across spectral slits. This means that the response patterns (and associated confusion matrices) are somewhat similar across the spectrum for these features. In contrast, the SyR is extremely low for Place, indicating virtually no overlap across frequency of the response patterns. This is one reason why the information gained by combining two or three slits is so great for Place – the errors associated with each spectral region are complementary and hence result in a huge gain in information transmitted when combined. This "Principle of Complementarity" is directly relevant to Fletcher's "Product of Error" (POE) rule and the Articulation Index (AI). What it shows is that the underlying basis for the AI and the POE rule in particular is the special nature of Place of Articulation information, and that studies of consonant recognition need to take into account the underlying phonetic features. Further details about this study can be found in the associated PDF file, which contains a copy of Christiansen and Greenberg (2012).

In papers yet to be published in archival journals (the data have been presented at conferences and published as book chapters – Christiansen et al., 2006; Greenberg and Christiansen, 2010) we show that the way in which consonants are processed is highly asymmetric and directional in terms of phonetic features. In "The Perceptual Flow of Phonetic Processing," consonant confusion matrices are analyzed for patterns of phonetic-feature decoding errors conditioned on whether other phonetic features are correctly or incorrectly decoded. The analyses demonstrate that the feature Voicing is correctly decoded most of the time and rarely depends on decoding Manner and Place correctly. However, Manner decoding does depend on getting Voicing decoded correctly; hence there is an asymmetric relationship between the decoding of these two features. Moreover, a similarly asymmetric relationship holds between Manner and Place decoding. Place decoding depends on Manner being decoding correctly, not vice versa. Moreover, the relationship between Place decoding and Voicing decoding is transitive – given the dependency of Manner decoding on Voicing, a comparable dependency on Place decoding (on Voicing decoding) is also observed. From these conditional probability patterns, it is proposed that they reflect a temporal flow of perceptual processing – Voicing is processed and decoded prior to Manner, which is processed prior to Place. It is likely that this "perceptual flow" is related to the distinctive pattern of phonetic-feature decoding observed in previous study. For Place to be accurately decoded, at least two of the three slits need to be present, and preferably all three. In contrast, both Manner and Voicing can be decoded well with one or two slits.

The third study examined the impact of low-pass filtering the modulation spectrum on consonant recognition and phonetic-feature decoding. The modulation spectrum of the Danish syllable stimuli was low-pass filtered between 24 Hz and 3 Hz, in octave steps (i.e., 24 Hz, 12 Hz, 6 Hz and 3 Hz). This low-pass modulation filtering was imposed on each spectral slit presented alone and in combination with the other slits. Not surprisingly, low-pass filtering had a decided impact on consonant recognition. The main question addressed in this study was whether the phonetic features were more affected by filtering in specific modulation-frequency regions. The answer is "yes," they were. Of particular interest is the finding that Place of Articulation is most sensitive to filtering in the 6 – 12 Hz region, while Manner and Voicing are more sensitive to filtering in the very low (< 6 Hz) and higher (>12 Hz) modulation frequency regions. In our view, this differential sensitive reflects the fact that Voicing and Manner reflect energy dynamics on a slow and fast time scale, somewhat irrespective of where the dynamics occur, while Place of Articulation is quite sensitive to the rate of the dynamics across most of the speech (acoustic-frequency) spectrum.

The three studies described in this section raise as many questions as they answer. In a follow-up study, Thomas Christiansen and his student, Maria del Pilar Ocaña Nuñez, have shown that consonant recognition and phonetic-feature decoding are very sensitive to vocalic context in the presence of band-stop noise. Their data imply there is a relatively long time constant for analyzing Place of Articulation information, which would be consistent with the earlier studies on the modulation spectral filtering and the perceptual flow of phonetic information. Thomas intends to follow up these issues in future perceptual studies.

DejaNets – A Theoretical Framework for Speech Recognition in Humans and Machines

Prologue – The Origins of DejaNets in the TEMPO model developed by Ghitza and Greenberg

The second project also grew out of previous work, this time with Oded Ghitza of Sensimentrics and Boston University. Oded had asked my advice at a research meeting (International Symposium on

Hearing) in 2006 about a proposal he was soon to submit to the AFOSR on word recognition. As a result, I became a consultant on that project the following year. From that early beginning, we began collaborating on an AFOSR project entitled "Decoding Speech Using Neural Rhythmicity." The main product of our collaboration was a study that formed the basis of a paper published in 2009 "On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence". The main result of that study is shown in the figure below (from Ghitza and Greenberg, 2009). There were two aspects of these data that are (in my opinion) particularly interesting. First, the insertion of silent gaps enhances intelligibility of otherwise difficult-to-comprehend, time-compressed speech when the gaps range between 20 ms and 120 ms. Usually, a distortion such as insertion of silence, degrades intelligibility (e.g., Huggins, 1975). In this instance, the gaps restored a large part of the intelligibility lost through the time compression. It's unusual for one distortion to counteract that of another. Obviously, something interesting was happening, but what?

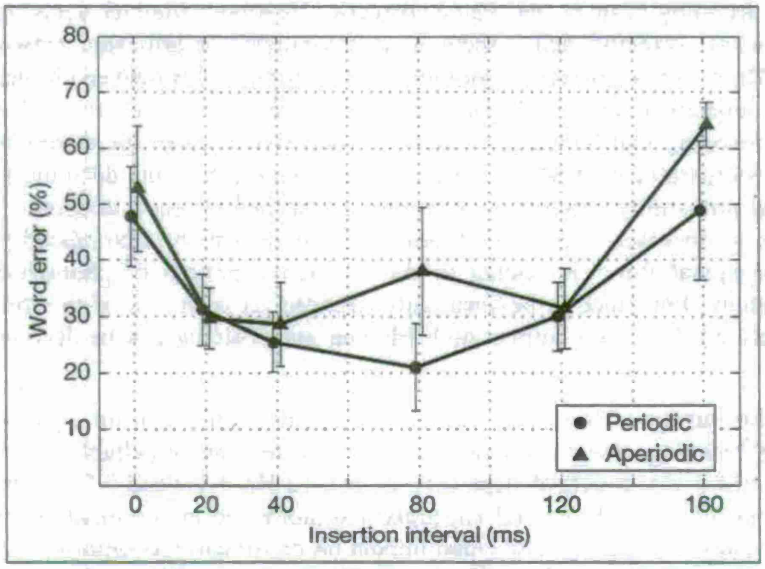


Fig. 2. Intelligibility of time-compressed speech as a function of the duration of inserted silence. Word error rate is plotted as a function of the silence-interval duration (error bars represent the standard deviation of the mean). Speech was time-compressed by a factor of 3. Speech segments are consecutive 40-ms-long intervals and are kept the same for all conditions.

One potential clue is the performance associated with the aperiodic silences. The 80-ms and 160-ms gaps have a largely negative impact in intelligibility. Could it be that that this is because some sort of quasi-periodic rhythmicity underlies speech decoding? This was the effect Oded and I were looking for, but as Willard Larkin pointed out, the result is somewhat weak.

We sketched out a much larger series of perceptual studies to perform in which different speech materials would be used (I created these, but left the project before they were deployed – see below), segmented at a range of lengths and interrupted by a broader range of silent intervals. These perceptual studies were designed to ascertain whether the patterns observed in the figure above were consistent across linguistic material and a broad range of interruptions and speech durations.

The figure below shows a cartoon sketch of the model (TEMPO) that might account for the intelligibility data in Ghitza and Greenberg's Figure 2. Template matching and memory (in a loose form) are important components. In particular, we hypothesized that there would be an important interaction between different types of brain rhythms that could underlie the brain's ability to decode speech.

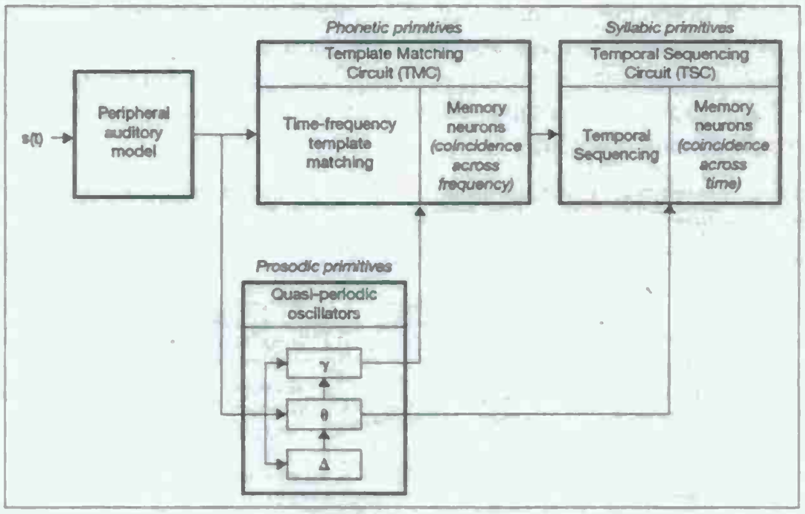

Phonetic primitives
Template Matching Circuit (TMC)
Syllabic primitives
Temporal Sequencing Circuit (TSC)
s(t)
Peripheral auditory model
Time-frequency template matching
Memory neurons (coincidence across frequency)
Temporal Sequencing
Memory neurons (coincidence across time)
Prosodic primitives
Quasi-periodic oscillators
γ
θ
Δ

Fig. 4. A block diagram of the TEMPO model. See text for details regarding the model.

It has been known for many years that large neural populations fire synchronously at slow rates. The first empirical evidence for such neural oscillations was produced in the 1930s (Han Berger recording "alpha" waves). Since that time a number of other oscillatory phenomena have been discovered. These range from "gamma" oscillations in the 30 Hz – 80 Hz range (also known as the "40-Hz" wave), all the way down to "delta" in the very low frequencies (< 4 Hz). In between these are "theta" in the 3 Hz – 10 Hz range and "beta" (10 Hz – 30 Hz) rhythms. Alpha oscillations (8 Hz – 12 Hz), the original brain waves, are now thought to be associated mainly with the relaxation state of the brain, and so are less important for speech processing.

As part of a project proposal for an AFOSR STTR solicitation, I began working out a more elaborate model than that included in the Ghitza and Greenberg (2009) paper. In particular, my focus was on how the neural oscillations could participate in the parsing of the speech signal, decomposing it into linguistic units such as phrases, words, syllables and segments. My intuition was that the *interaction* among the different oscillators could play an important role in the parsing, particularly given the close match between the time course of these oscillators and the temporal properties of speech. Delta oscillations are similar in duration to complex words and phrases; theta oscillators are closely matched in time to syllables and short words, beta oscillators might correspond to phonetic segments and gamma oscillators to phonetic features and primitives.

I was also struck by the fact that these different time courses are reflected in the modulation spectrum of speech, which had formed the focus of my research over the previous dozen years. I proposed to Oded that the low-frequency modulations observed in the speech signal could act as "triggers" effectively setting off oscillatory activity in the cerebral cortex. The STTR proposal was supposed to use this concept as a way of parsing the speech signal into linguistic units that could eventually be used to decode phonetic and lexical units in automatic speech recognition.

For a number of reasons, the STTR proposal was never submitted. However, before this decision was made I had developed a rough model of how the speech signal could be parsed using neural oscillators, as well as performed a lot of background research. For reasons explained below, this background research expanded dramatically over the next two years and is the reason why an archival paper has yet to be completed.

The project's program manager, Willard Larkin gave me permission to change the focus of the AFOSR speech synthesis project to focus on brain rhythms and speech perception (if my understanding is correct) any other related topic I believed was important for understanding how the brain understands spoken language. I am profoundly grateful for this permission because it enabled me to develop a theoretical framework – DejaNets – that accounts for many (if not almost all) properties of spoken language. It represents a significant extension of what Oded and I had done on the neural rhythmicity project and offers the prospect of changing the way in which speech research is performed and such knowledge used in science and technology.

Before describing DejaNets, I would like to discuss certain aspects of the TEMPO model, as represented in a recent paper by Oded, "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm" as shown in the figure below.

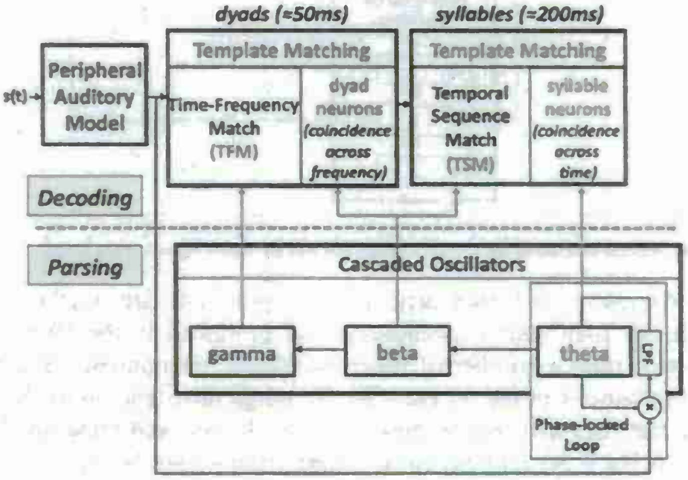

dyads (≈50ms)
syllables (≈200ms)
Template Matching
Template Matching
s(t)
Peripheral Auditory Model
Time-Frequency Match (TFM)
dyad neurons (coincidence across frequency)
Temporal Sequence Match (TSM)
syllable neurons (coincidence across time)
Decoding
Parsing
Cascaded Oscillators
gamma
beta
theta
LPF
Phase-locked Loop

FIGURE 1 | A block diagram of theTempo model. It comprises lower and upper paths that process the sensory stream generated by a model of the auditory periphery. The lower path extracts *parsing* information, which controls the *decoding* process performed in the upper path. The parsing is expressed in the form of an *internal clock-like mechanism*, realized as an array of cascaded oscillators locked to the input syllabic rhythm; the frequencies and relative phases of the oscillations determine the processing time frames that control the decoding process. See text for details.

This figure is a slight elaboration of Figure 2 from Ghitza and Greenberg. The terms "Decoding" and "Parsing" have been added (these were in the STTR proposal that was never sent). Also added is a "phase-locked loop" intended to provide a mechanism for describing how acoustic modulations are transformed into theta oscillations. "Dyad neurons" have been added to make the shorter time frame more explicit linguistically. Perhaps the most significant addition is the term "cascaded oscillators," which is the term Oded uses for what were referred to as an "oscillatory hierarchy" in the aborted STTR proposal. In the proposal, a prominent role was to be given to delta rhythms because of their organizational role in resetting the phase of theta oscillations (Schroeder and Lakatos, 2009). Oded argued in his recent article that delta oscillations lay outside the scope of his current model, so omitted them from the TEMPO framework. One other addition was made, which is shown in Ghitza (2011): Figure 2, shown below. The difference in modulation phase (as manifest in the specific timing of peaks and valleys in the envelope) across the tonotopic frequency array is clearly shown. Surprising, Oded doesn't address how these phase differences are reconciled in TEMPO. I had raised this crucial issue of tonotopic organization and auditory filtering with Oded shortly before our STTR proposal was scuttled, but this issue was labeled as "trivial and obvious." As will be shown later in this report, the modulation phase disparity across the acoustic frequency spectrum is likely a key component of how the auditory system triggers into the memory system via gamma, beta and theta oscillations.

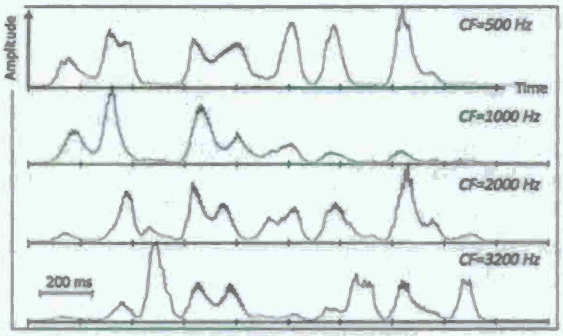


FIGURE 2 | Cochlear envelopes in terms of a simulated inner hair cell response, low-pass filtered to 50 Hz, at four characteristic frequencies. The input is the semantically unpredictable sentence "The ripe style heard their spades," naturally spoken by a male speaker (from Ghitza and Greenberg, 2009). The duration of the signal is roughly 2 s (ten 200-ms long frames). The rate of the envelope fluctuations is about 5 peaks per second. Low-frequency cochlear channels mainly reflect the presence of vowels and nasals, and high frequency channels mainly reflect fricatives and stop-consonants. The PLL (see **Figure 1**) is locked to the temporal fluctuations of the cortical auditory representation of the speech signal (e.g., the modulation spectrum), which is related to the cochlear response.

The TEMPO model focuses on decoding semantically anomalous sentences, none of which are ever likely to be encountered in the real world. This restriction was imposed to neutralize semantic factors and as a way to focus entirely on phonetic and syllabic decoding devoid of meaning. This narrow focus makes it difficult to link the perceptual data displayed in Ghitza and Greenberg (2009) Fig 2 to spoken language processing, except in the most artificial way (a problem present for ca. 90% of speech perception studies, including those described in the initial part of this research report ☺). The problem for the TEMPO model is its focus on theta and other brain oscillations. Theta oscillations in particular have been linked to memory processes. The sort of memory invoked for decoding words in semantically unusual contexts may differ significantly from that in more realistic conditions. So, the question naturally arises "what is meant by memory" in the TEMPO model? We return to this issue in the discussion of DejaNets below.

There are other problems with the TEMPO model and the data that underlie it. First, the quality of the perceptual data is suspect. They were collected in an informal way by Oded not in a controlled laboratory setting. At least one of the subjects was a member of his family and may have known what the underlying logic of the experiment was. Moreover, half of the perceptual data were omitted in describing the TEMPO model. The figure below shows the data in Ghitza (2011). Note the absence of the aperiodically silence data from the original Ghitza and Greenberg (2009) study.

What happened to the other data? Perhaps they were too difficult to model with the rigid, clockwork oscillatory framework that had become TEMPO. Although the data in the Ghitza (2011) show the effect of inserting silent gaps most clearly, the aperiodic silence data are intriguing precisely because they were so similar for the 20-ms, 40-ms and 120-ms points and so different for the 80-ms and 160-ms points. Given the clockwork nature of TEMPO, the aperiodicity of the gaps is a crucial test of the model. The decline in intelligibility (of nearly 20%) at the 80-ms point for the aperiodic condition is striking. Why this disparity?

One potential problem with both sets of data (associated with the periodic and aperiodic gaps) is the confounding of silent gap duration and sentential material. The same linguistic material was presented to all subjects for a given condition. No attempt was made to dissociate the sentences from the silence conditions. Hence, the performance for any given condition could reflect, at least in part, the difficulty (or ease) of decoding. This may explain why the intelligibility associated with the aperiodic, 80-ms and 160-ms conditions are so much poorer than the 40-ms and 120-ms aperiodic conditions. On the other hand, the intelligibility disparity may reflect bona fide mechanisms. Until the experiment is redone this will remain an open question. I hope Oded redoes the study, as the data are potentially important for developing models of spoken-language processing. The original design of the study came from a question Oded posed to me several years ago: "how would one test the possibility that brain rhythms underlie speech perception?" I suggested using a variation of a paradigm that Huggins (1975) had used to ascertain the temporal parameters associated with working memory for decoding speech. The crucial suggestion (in my opinion) was made by Willard Larkin to time-compress the speech as a way of stressing the listeners. Combining the Huggins paradigm with time compression allowed the impact of decoding time to be more clearly demonstrated than we had been able to show using a simpler design.

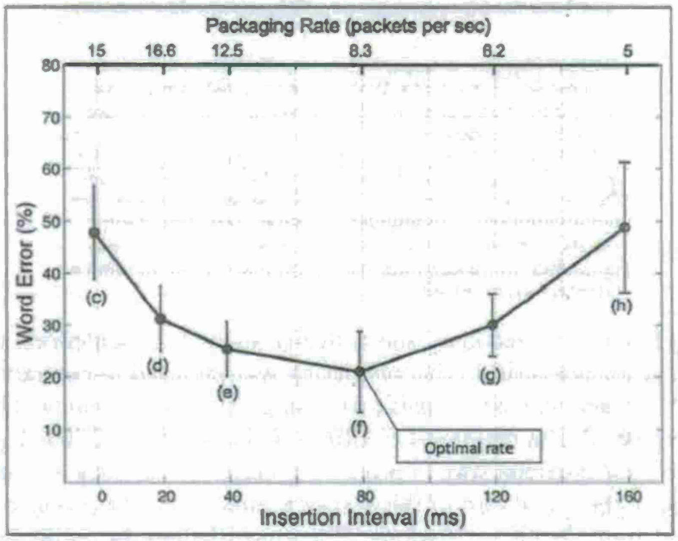
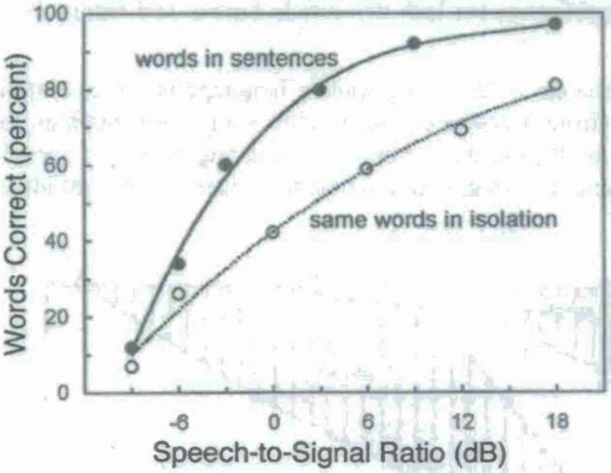


FIGURE 3 | Intelligibility of time-compressed speech with inserted silent gaps (from Ghitza and Greenberg, 2009). The signal conditions are labeled (c–h), in correspondence with the labeling in **Figure 4** and **Table 1**. Word error rate is plotted as a function of gap duration or, equivalently, packaging rate. Speech was time-compressed by a factor of 3. Acoustic intervals were consecutive 40-ms long speech intervals, kept the same for all conditions. Without insertions performance is poor (>50% word error rate). Counter-intuitively, the insertion of gaps improves performance, resulting in a U-shaped performance curve. The lowest word error rate (i.e., highest intelligibility) occurs when the gap was 80-ms long (or, equivalently, at the rate of 8.3 packets/s), down to ca. 20% (the "optimal" rate). (The intelligibility of uncompressed speech, and speech compressed by a factor of 2, is high, with error rate <2%.)

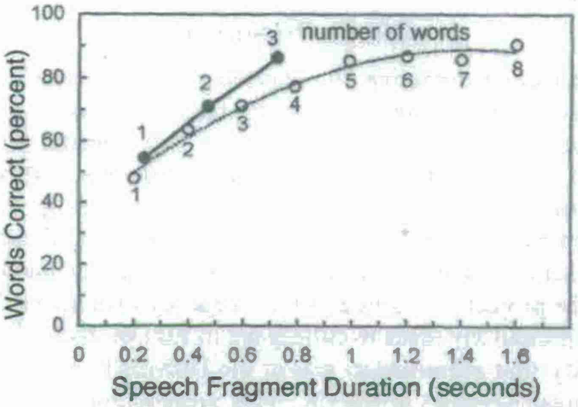
TEMPO, as described in Ghitza (2011) is not really a model in the normal sense of the term. It's more of a phenomenological perspective with little computational support, only some graphs and pictures to illustrate the general concept. Moreover, it is derivative of a model proposed by John Lisman many years ago (but not cited in the article) in which the phase of theta oscillations in the hippocampus could be used as a form of short-term memory encoding (this work influenced David Poeppel's research). Moreover, TEMPO is too rigid in its clockwork structure, whereby each oscillatory cycle is composed of an integral number of oscillatory cycles from a lower-level oscillation (e.g., each theta cycle being composed of 4 beta cycles, etc.). Schroeder and Lakatos' research in auditory and visual cortices (of rhesus monkeys) is inconsistent with this rigid framework (as delta activity can reset the phase of theta oscillations, implying that phase reset could be an important way for the top-down flow of information to proceed). TEMPO is essentially a bottom-up, mechanistic and largely deterministic framework that is inconsistent with most of what is known about spoken language (and human behavior in general).

DejaNets – A Theoretical Framework for Understanding Spoken Language

The Russian linguist, Roman Jakobson, once wrote “We speak, to be heard, to be understood” (Jakobson, Fant and Halle, 1962). Jakobson was articulating what many who perform speech research fail to understand –the act of listening is primarily a ***search for meaning***. Speech sounds and vocal gestures are only a (very limited) means towards understanding and using this knowledge behaviorally. The figures below illustrate the importance of semantic (and syntactic) context for intelligibility.



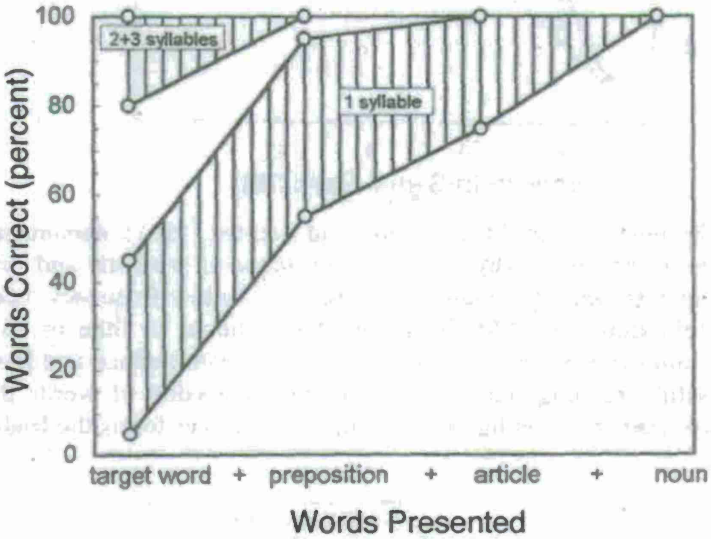
The figure above, originally published by Miller, Heise and Lichten (1951), demonstrates that the ***same acoustic signal*** is more easily decodable (by far) in the presence of semantic and syntactic context. If speech parsing and decoding were ***merely*** a matter of using acoustic low-frequency modulations to evoke theta, beta and gamma oscillations, as TEMPO implies, there should be little or no difference in the intelligibility curves. This point is reinforced by the figure below from Pollack and Pickett (1962). Their study showed that successful decoding requires a minimum succession of words that last at least 1 second. Why should reliable speech decoding take so long? The answer forms the basis of the remainder of this project report.



The short (but incomplete) answer is that delta oscillations are likely at work in guiding the analysis and interpretation of the speech signal. They reflect very long-range communication across many different cortical and hippocampal (and probably thalamic) regions, and are the physiological manifestation of extensive cross-modal, poly-sensory integration that is largely governed by memory processes. This brief answer is not really an explanation but only a supposition that connects very low-frequency oscillations with speech decoding (and by implication, speech understanding).

Theta oscillations (3-10 Hz) are thought to reflect primarily hippocampal-cortical interactions because they are most visible in certain parts of the hippocampus during tasks heavily dependent on short-term memory. Their presence is consistent with lesion as well as physiological studies demonstrating a close relationship between short-term memory and this structure of the brain (i.e., hippocampus). Theta activity is observed in other brain regions, but it is unclear whether it occurs when extensive interpretation of the sensory signal is required. It is possible that delta activity is a signature of "deep" thinking and that its absence in many experimental studies is more a reflection of the experimental design and behavioral task than a reflection of delta's significance (or lack thereof) in human and animal behavior. We return to this issue below.

More evidence that the process of decoding spoken language is not as simple as TEMPO implies is shown in the figure below (from Grosjean, 1985). This study demonstrates that syllable structure is extremely important in word intelligibility, and that there is a huge range of variation in intrinsic decoding ability, depending on the phonetic, syllabic and syntactic properties of individual words. We revisit this issue below.



Based on the work I had done preparing the aborted STTR proposal, I continued to reformulate TEMPO into a general theoretical framework, which eventually became DejaNets ("deja" already experienced [memory], "nets" shorthand for hippocampal and cortical networks). TEMPO was an important jumping off point for this later work, and the discussions with Oded over the years were important in forming my own views about the role of brain rhythms in spoken language. However, the background research I performed for the STTR proposal had already convinced me that TEMPO was both incomplete and in certain respects wrong-headed. Now that I was no longer actively engaged in the theoretical development of the project, I was free to formulate my own theoretical perspective. Beginning in October, 2009, I began to present my ideas to colleagues in Europe, the U.S. and Asia in order to solicit their advice and feedback. My first presentation was at the University of Sheffield (Speech and Hearing group, Department of Computer Science) in the UK. This presentation was followed by presentations at the Technical University of Denmark (Centre for Applied Hearing Research), Ludwigs Maximilian Universität, Munich (Speech and Phonetics) and at several Japanese universities (Sophia in Tokyo (Electrical Engineering), Doshisha (Neuroscience) in Kyoto, and Wakayama (Electrical Engineering)). The reaction was mixed but encouraging. The general consensus was that DejaNets was novel and potentially explanatory. Some audiences were more receptive than others, but there were always some in the audiences who were genuinely enthusiastic. Some (usually older researchers) were more circumspect or skeptical (particularly at Sheffield and Munich).

The main points I made in these early presentations is that "time" and "memory" are extremely important for decoding and understanding speech. I made a general case for memory circuits operating at different time intervals controlling the interpretation of inherently ambiguous sensory signals. I also pointed out the importance of sensory integration such as the McGurk effect and other audio-visual interactions demonstrate. But I was also interested in formulating a general theory of spoken language that could account of dozens of otherwise mysterious properties of speech. In this effort, I believe DejaNets is truly distinctive. It aims to be a theory of nearly everything (linguistic and cortical) and could also account for many other aspects of human (and probably) animal behavior.

I tried finding an academic partner who could perform specific experiments to test the theory. Over the course of two years I have spoken with David Poeppel at NYU and have given a presentation to his group. Despite many discussions and expressions of interest, nothing came of this (I recently learned that Oded has an active research collaboration with David, so perhaps this is why David has been unwilling to also collaborate with me). Sue Denham of Plymouth University in the UK was sufficiently interested that she tried to find a student to work on the project. To our surprise, no qualified student was found. I spoke with Lee Miller at UC Davis about the possibility of collaboration and gave an informal presentation outlining the theory. It was clear from his reaction and that of his students that their interests lay elsewhere. I spoke with a just-finished PhD student in Munich, whose background was in speech and who came highly recommended. Again, his interests lay elsewhere. I also tried to persuade my colleagues at DTU in Denmark, but no one in that group was interested in collaborating on this project either. One could interpret this lack of finding a research collaborator in either a negative or positive light. Perhaps the theory is just not sufficiently interesting to attract a collaboration or perhaps the ideas are too ahead of their time? Or perhaps no one was interested in collaborating with me personally. It's difficult to know.

Eventually, I concluded that it made the most sense to focus on theory development and not try to perform perceptual studies during the remainder of the project or to waste further time and money looking for collaborators.

What follows is an outline of the theoretical framework associated with DejaNets. A full description would require more than 50 pages, so only a brief distillation is included in this report. Further information is available upon request (and will be included in the BBS manuscript submitted later this year). I delivered an invited talk in a special session on "Invariance" at the Acoustical Society of America Meeting in San Diego (Nov. 1, 2011). This abstract serves as a useful way to introduce the DejaNets framework:

Speak, Memory–Wherefore Art Thou Invariance?

Spoken language is highly variable, reflecting factors of environmental (e.g., acoustic-background noise, reverberation), linguistic (e.g., speaking-style) and idiosyncratic (e.g., voice-quality) origin. Despite such variability listeners rarely experience difficulty understanding speech. What brain mechanisms underlie this perceptual resilience, and where does the invariance reside (if anywhere) that enables the signal to be reliably decoded and understood? A theoretical framework – DejaNets – is described for how the brain may go from "sound to meaning." Key is speech representations in memory, crucial for the parsing, analysis and interpretation of sensory signals. The acoustic waveform is viewed as inherently ambiguous, its interpretation dependent on combining data streams, some sensory (e.g., visual-speech cues), others internal, derived from memory and knowledge schema. This interpretative process is mediated by a hierarchical network of neural oscillators spanning a broad range of time constants (ca. 15 ms–2,000 ms), consistent with the temporal structure of spoken language. They reflect data-fetching, parsing and pattern-matching involved in decoding and interpreting the speech signal. DejaNets accounts for many (otherwise) paradoxical and mysterious properties of spoken language including categorical perception, the McGurk effect, phonemic restoration, semantic context and robustness/sensitivity to variation in pronunciation, speaking rate and the ambient acoustic environment. [Supported by AFOSR]

The basic principle underlying DejaNets is that perception relies on an intricate coordination between the sensory and memory (and probably motor) systems. Although this may seem obvious, most models of speech perception either ignore the memory system entirely or (in the case of TEMPO) give it short shrift, at best positing a nebulous form of pattern matching without describing how such an operation would proceed nor describing its implications for how such information is structured and processed.

Within the present framework, the memory systems are crucial for understanding how the sensory systems are organized; moreover, they provide important insights into why speech is structured in the way it is.

Neither the spectral nor temporal properties of speech are easily understood within the "traditional" articulatory-phonetic framework (hereafter, ArtPhon). They are assumed to reflect primarily biomechanical constraints of the vocal apparatus. The tongue, lips and jaw can move only so fast. The spectral changes associated with the resonances (formants) of the vocal tract reflect the inertial properties of tongue motion. Syllables are, within this framework, merely reflections of articulatory gestures. Their phonetic composition and variability are due to the necessity of distinguishing different linguistic elements (i.e., the phonetic structure of speech is largely arbitrary, subject to biomechanical constraints). It was such logic that ultimately led to the "motor theory" of speech, which posits that listeners decode speech by back-computing the articulatory gestures that produced the acoustic signal.

No model developed within the ArtPhon framework can adequately explain the following speech properties and phenomena:

(1) Perceptual invariance in the presence of enormous sensory variability

(2) Dynamic sensory input resulting in the perception of discrete phonetic elements

(3) Categorical perception, where this is apparent stability over a large range, combined with abrupt perceptual shifts

(4) Insensitivity to intrusions into the speech signal under many conditions ("phonemic restoration")

(5) Ability to decode speech when much of the acoustic signal is either missing or highly distorted (in either time or spectral frequency)

(6) Ability to decode speech despite a large variation in speaking rate

(7) Sentential context is often required to decode individual words

(8) Noise improving speech decoding under certain conditions ("perceptual induction")

(9) Delayed auditory feedback exerting a destructive effect on speaking over certain feedback delay intervals

(10) Visual speech cues improving intelligibility under many (but not all) conditions

(11) Semantic context dramatically improving the ability to decode speech, particularly in noisy conditions (or among the hearing impaired)

DejaNets can account for these phenomena (and many more). It does so by "working back" from the pattern recognition operations involved with memory to the cortical mechanisms associated with this recognition (cortical and hippocampal oscillations) to the structure of the sensory signals (low- and mid-frequency modulations) distributed across the auditory tonotopic axis.

Although a full exposition of the theory lies outside the scope of this project report, a few examples (and figures) are provided to convey a sense of how DejaNets operates and its potential explanatory power.

Perceptual Invariance

Cortical oscillations are inherently inertial. Once a sensory signal triggers a deja (i.e., distributed memory) network, its "contents" are essentially "frozen" in time. In other words, the sensory signal, no matter how dynamic its composition, is linked to an oscillatory event that is essentially static in terms of conscious decomposition. Because each oscillatory event is associated with some form of memory (i.e., a deja), it provides a mechanism for highly dynamic input, such as formant trajectories in speech, to be "perceived" as discrete entities. In some sense, this perception of speech as a sequence of words, syllables and phonetic segments is an illusion resulting from an ex-post-facto analysis for linguistic communication (such as speaking or writing). The variability of the speech signal is far too great and the conversion from dynamic input to small abstract units (such as phone, syllable or word) too time-consuming for this sort of process to occur "on the fly" in real time. It is more likely that the incoming speech signal is "stored" in some form of compressed "raw" form. This is why familiar voices are more intelligible than ones not previously encountered (shown by Goldinger, 1997; 2004). This is also why most of us can imitate a wide range of accents and speaking styles. But speech is normally used for the transmission of "meaning," so that there is no conscious awareness of more abstract units during speech reception unless one's attention is specifically directed to this level (moreover, the conscious conversion to phonetic segments and syllables is very slow, and usually requires some phonetic training). Most of the time, the speech signal is "transformed" directly into meaning without conscious awareness of its underlying constituents. This is because the speech signal is not directly decoded into phones, syllables and words (as the "conventional" models assume) but rather acts to "trigger" special memory ("deja") networks that "seek" a meaningful interpretation. Usually, these triggers come from the acoustic signal, but they can also derive from visual cues associated with the movement of the lips, jaw and tongue during speaking. These extra-acoustic triggers can enhance intelligibility (and meaning) dramatically (e.g., Greenberg and Arai, 2004), and where this information adds super-linearly in a manner comparable to place-of-articulation information in the acoustic signal. Deja nets can also be triggered by extra-sensory information, including knowledge of the semantic and social situation associated with the spoken discourse. It is the interaction between the external (i.e., sensory) and internal (i.e., memory) systems that allow speech to be so effortlessly understood most of the time. Degradation of either seriously compromises the ability to understand, either through hearing impairment, acoustic interference or cognitive memory problems confronting the elderly and victims of brain trauma.

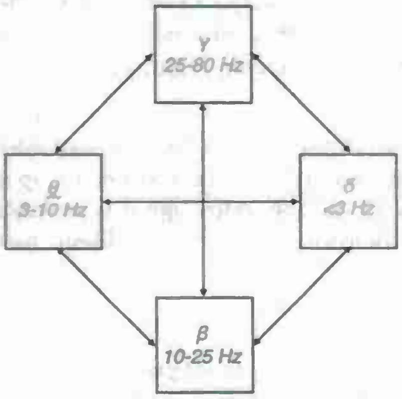
So, where does the "invariance" lie? Most often in the meaning associated with the speech signal, not in the signal (or its auditory representation)? Decomposition of speech into "units" associated with phones, syllables and words is largely an illusion. Although it is possible to pinpoint different parts of the speech signal and link these to such linguistic units, it is unlikely that the brain decomposes the input signal in this way.

Speech Robustness

One of the most remarkable properties of spoken language is its ability to be understood across a vast range of conditions – very low signal-to-noise ratios, highly reverberant environments, variability in speaking rate, phonetic and prosodic realization, and contextual information. In the Ghitza and Greenberg (2009) study, the distortions imposed – radical, uniform time compression and insertion of silent gaps – exerted a differential effect on intelligibility, depending on the length of the silent gap. The variation in intelligibility associated with gap length can be interpreted in a several ways. Huggins (1975) posited a working memory buffer of ca. 200 ms in his original "gap-bridging" study upon which the Ghitza and Greenberg (2009) paradigm is based. The problem with Huggins' hypothesis is that it doesn't explain why intelligibility is so poor unless the gap is 40 ms or longer. However, his memory-buffer model is consistent with intelligibility's decline for gaps greater than 120 ms.

To explain the U-shaped intelligibility curve, Oded and I suggested that cortical oscillators with certain time constants could be involved. Implicit in the TEMPO model is the assumption that pattern matching of input with stored "templates" is a vital part of the speech recognition process. Time compression degrades the pattern-matching operation and hence the ability to match the input signal with "stored templates" residing somewhere in memory. TEMPO's attraction is that it provides an intuitive way of understanding why time compression has little impact on intelligibility up to a certain limit (ca. 200%) and why compression ratios much greater than 200% are devastating. Over a dynamic range of four, from 50% compression (200% time expansion) to 200% compression, speech appears relatively immune to temporal distortion. This range roughly matches that of theta oscillation frequency (3–10 Hz).

However, there are important properties of theta oscillations (and other brain rhythms) that don't meld well with the TEMPO model described in Ghitza (2011). Wolf Singer performed a disservice to neuroscience by referring to gamma oscillations as the "40-Hz oscillation." This name gave the misleading impression that brain rhythms occur at a single frequency – they do not. What is key about cortical oscillations is the frequencies over which they vary. Theta ranges *between* 3 and 10 Hz, it does not oscillate just at 6 Hz or 4 Hz, as some researchers imply. It is this flexibility in the theta frequency that's key for understanding its likely role in neural information processing. The other brain rhythms also vary in their frequencies and by comparable amounts (on a logarithmic scale). TEMPO implies that there is a rigid, quasi-harmonic relationship between oscillatory frequencies at different time scales (i.e., theta, beta, gamma). This is unlikely to be the case. One of the strongest pieces of evidence against this "harmonic" model of brain rhythms is the work by Schroeder and Lakatos (2009) who demonstrate that the relationship between theta and delta oscillations is highly dynamic, and that the phase of the former can be "reset" by the latter. This is precisely what would be expected if there were a hierarchical structure to the different oscillatory time scales, for it implies that longer time-scale rhythms (e.g., delta) could govern the phase dynamics (and hence timing) of shorter, higher-frequency oscillations such as theta. This may be why Oded refers to TEMPO as a "cascaded" series of oscillators. However, the analogy is not particularly apt. This is because the term cascade (and his figure 1 from Ghitza, 2011) imply a uni-directional flow of control (from theta on down to beta and gamma). It is more likely that the relationship among oscillatory time scales is multi-direction such as shown below:



In this scheme, higher-frequency oscillations can impact lower-frequency ones (and vice versa). Such interactional flexibility is required to account for the ability of short-time-scale phenomena (such as a single phonetic segment or phonetic features) to impact the interpretation of the longer-time-scale phenomena (at the word or phrase level). For example, the meaning of the word "look" can be modified dramatically by the simple addition of a single segment, such as [s] (looks), [t] (looked) or even [ae] (lack). The ability of very short-time-scale phenomena to impact the interpretation and meaning of longer-time scale entities cannot easily be accommodated within a cascaded oscillatory framework (even with a phase-locked loop).

Another significant challenge to TEMPO and the Ghitza and Greenberg (2009) data is the impact of non-uniform time compression on intelligibility. Covell, Withgott and Slaney (1998) developed a special-purpose time-compression system, Mach1, which compresses constituents of the syllable differentially (I thank Malcolm Slaney for bring this study to my attention). In their system, vowels of stressed syllables (i.e., containing the most energy) were compressed *the least*, while consonants in unstressed syllables were compressed the most. Under such circumstances, intelligibility was much better for the Mach1 system relative to uniform compression:

Table 1: Comprehension-rate differences between Mach1-compressed and linearly compressed speech, by test section. Significance levels for each section are also shown.

Section type	Comprehension rates (in percentage points)	
	Average	Difference (Mach1 – linear)
short dialogs	70	31.0 ($t'_{52}=8.60\ p<0.001$)
	82	14.8 ($t'_{52}=4.13\ p<0.001$)
	76	24.3 ($t'_{52}=6.79\ p<0.001$)
long dialogs	79	5.4 ($t'_{52}=1.50\ p=0.702$)
monologs	81	10.0 ($t'_{52}=2.79\ p=0.036$)

In effect, these results imply that time's importance varies as a function of the amount of meaningful information contained at different points in an utterance. In other words, the effect of time demonstrated in Ghitza and Greenberg (2009) cannot realistically be divorced from the semantic interpretation of the signal. A demonstration of the Mach1 system is available at:

http://www.mangolassi.org/covell/1997-061/

The Mach1 time-compression system also provides some important insight as to why speech is temporally structured in the way it is. One of the great paradoxes of speech perception is that consonants are widely acknowledged to convey most of the linguistically discriminative information distinguishing words and hence crucial for meaning, yet vowels appear to be crucial for intelligibility, particularly in noise (inferred from some recent research by Pierre Divenyi and Diane Kewley-Port). How can it be that the less-informative parts of the signal are more important for understanding speech?

Vowels in stressed syllables are considerably longer than in their unstressed counterparts. Stressed syllables are also linguistically more informative. Could it be that the duration of vowels is associated with the time required for the brain to appropriately process and interpret the information contained within the entire syllable? And could it be that unstressed consonants can be compressed to a greater degree than stressed vowels because the information with which they are associated is largely predictable (i.e., redundant) from that associated with stressed syllables? And could it be that the reason why even uniformly compressed speech is intelligible for compression ratios of less than 200% is that the deja triggers are effective over a range of time scales because they are based on acceleration (double delta) parameters, rather than absolute or velocity-sensitive (delta) ones? Consistent with this conjecture is the finding that auditory cortical neurons act like accelerometers under many stimulus conditions (Heil, 1999). In other words, auditory cortical neurons (and probably other sensory cortical cells) are likely to act as "triggers" over a wide range of listening conditions, and may explain in part the brain's ability to understand speech in background noise and other forms of "distortion."

The Importance of Time

The Mach1 system provides an interesting way to gain insight into the neural mechanisms underlying the processing and (ultimate) understanding of speech. It is able to compress speech differentially depending on its short-term energy characteristics, and therefore could eventually deployed in hearing aids, cochlear implants and other devices where intelligibility is severely compromised. I recently presented some of the ideas below to the Starkey Hearing Research Center in Berkeley, CA. There was a lot of interest in the approach outlined below and there is a possibility of future collaboration.

The Mach1 system was designed for efficient time compression (ca. 250%) capable of providing reasonable intelligibility. However, Mach1 (or equivalent differential time-compression system) could also be deployed to provide *real-time* time compression of the less important components of the speech signal and *expansion* of the information-rich components that are crucial for intelligibility in low-signal-to-noise-ratio and highly reverberant conditions. In other words, the global (i.e., phrasal) time compression ratio would be 100% (i.e., no compression on the global time scale), but with compression and expansion distributed across the linguistic phrase, sentence and even syllable and word. Informal experiments I've performed recently suggest that the speech is likely to be more comprehensible at low SNRs than untreated material. Such a system could help not only the hearing-challenged, but also those (such as the US military) who operate is less-than-ideal acoustic conditions.

Another area where differential time compression could be deployed is in language learning, both native and foreign. By being able to emphasize certain syllables, words or even segments relative to others in an utterance, it might be possible highlight specific regions where the student is having problems and make it easier for him/her to understand the nature of the problem and correct it.

The Mach1 study by Covell et al. shows that there is another way to study brain rhythms and the importance of time, other than using silent gaps. Gaps produce distortion in the speech signal and are highly unnatural. By differentially time-compressing the speech signal, it should be possible to distinguish the memory trigger component of the process from the coordination and interpretation component. Uniform time compression makes it difficult to distinguish because because all parts of the speech signal are time-compressed equally. The "solution" that Oded and I discussed to overcome this problem was to present differential uniform compression ratios and insert variable lengths of silent gaps. This procedure doesn't truly solve the problem because the gaps are currently inserted at quasi-random intervals, dictated by the paradigm in which the gaps occur every 40 ms of time-compressed signal – regardless of the phonetic and prosodic composition of the speech before and after the gap. The Mach1 approach elegantly solves this conundrum by allowing certain phonetic and syllable constituents to be more highly compressed than others. Because the Mach1 system is quite old (late 1990s) and none of the study's authors are doing this sort of research anymore, a more practical approach to emulating the differential time compression is through the STRAIGHT vocoder system developed by Hideki Kawahara and his colleagues. STRAIGHT has precisely the properties required to differentially time-compress the signal (and in real time, thanks to an implementation created by Hideki Banno). All that would be required is an "executive" supervising STRAIGHT's compression and expansion based on principles similar to what Covell et al. used in Mach1, but adapted to the specific application (where the specific time-compression ratios might differ).

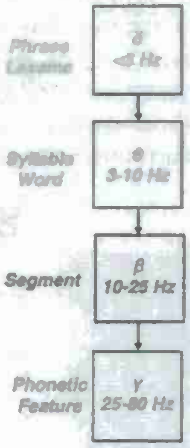
At the Acoustical Society of America meeting in San Diego in November, 2011, I gave a presentation describing DejaNets. Afterwards, Diane Kewley-Port mentioned to me that some of her recent speech perception research may be related. In essence, she and her colleagues have found that the vocalic portion of the syllable is far more important for comprehending speech than would be predicted from purely linguistic considerations (this is my interpretation of her results, but she did not disagree with it). As mentioned, the most plausible interpretation of Diane's study is that some neural process(es) crucial for speech comprehension require a certain amount of time to complete, and when that time is shortened, the brain's ability to "put it all together" degrades. This observation is of potential significance not only for the hearing-impaired or listening under low SNR conditions, but also for the cognitively challenged, particularly the elderly who appear to be particularly susceptible to speech-processing difficulties. I hope to explore the potential of this technology some time in the near future with a variety of academic and commercial organizations.

Time is relevant to cortical oscillations because of the range of time scales over which occur. The fast oscillations, gamma, almost surely reflect mainly local integration and processing, perhaps within a single cortical column or adjacent ones. This time scale is too short for the large-scale integration associated with memory processes and decision-making. However, gamma is probably critical in the memory-triggering process. The gamma patterns resulting from incoming sensory signals need to be configured just so for appropriate Deja networks to become active. This is probably why hearing impairment has such a devastating impact on the ability to understand – the sensory pattern of activation at the gamma time scale is insufficient to trigger the appropriate theta time-scale activity.

In the latest incarnation of TEMPO, the delta rhythms are omitted. This is a shame, as there is increasing evidence that long-range integration of cortical activity, involving the hippocampus and prefrontal cortex, is crucial for successful decision-making and task performance. By focusing primarily on theta and beta oscillations, TEMPO relegates speech processing to largely non-semantic material. Roman Jakobson's dictum "We speak in order to be heard and need to be heard in order to be understood" has effectively been cast aside. Speech for designed largely for conveying meaning, not to be micro-analyzed as part of a reductionist research project.

The Relation Between Low-Frequency Modulations and Cortical Oscillations

The temporal structure of speech bares an uncanny similarity to the time scales of cortical oscillations, as shown below. This correspondence is probably not coincidental. However, the precise relationship between linguistic units and oscillatory rhythms has yet to be worked out. Some, such as Oded and David Poeppel reasoned that speaking rate would be reflected in the frequency of theta oscillations. There thinking may have been based on one of the points I made in a presentation several years ago that the low-frequency modulations that are highly correlated with syllables could serve as a data-rate calibration mechanism so that the theta oscillations would adjust in some way to the variation in speaking rate. However, I did not intend to imply that there would be a one-to-one relationship between the low-frequency, modulation spectrum and theta oscillations. There are several reasons why. One is that there is rarely a one-to-one mapping between sensory and cognitive processing. Second, the modulation spectrum varies across the acoustic frequency (tonotopic) axis, so that it would be difficult, if not impossible, to derive a single modulation frequency from the speech waveform. There are multiple estimates of the modulation frequency, and this variation across the tonotopic axis likely plays an important role in the trigger mechanism for gamma and theta oscillatory activity.



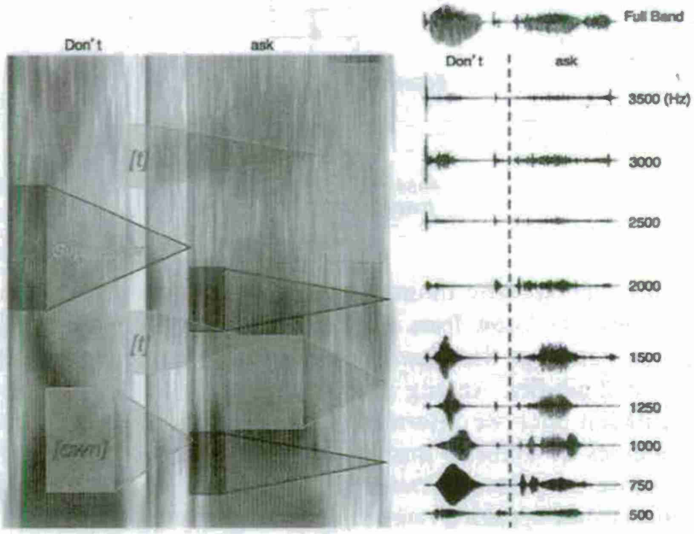
The modulation phase (i.e., the specific timing of the peaks and valleys of the energy fluctuations) varies across acoustic frequency. We know from the study that Takayuki Arai and I (e.g., Greenberg and Arai, 2004) performed several years ago that changing the modulation phase across the acoustic spectrum has devastating effects on intelligibility. Among other things, this phase distortion greatly impacts both manner- and place-of-articulation phonetic information, which is required for both segment and lexical discriminability. So, it comes as no surprise that an attempt by Oded, David and some of the latter's students failed to show a close correlated between the low-frequency modulation spectrum and theta oscillations in human listeners as the speaking rate varied (or was manipulated).

In the DejaNet framework, certain properties in the speech signal serve as oscillatory triggers. They need not be perfectly correlated with the global (i.e., full spectrum) waveform modulation. And often they are not. Instead, what is probably important is the phase (i.e., time) disparity between modulation peaks (or their onsets) across tonotopic frequency. That phase disparity serves as a key signature of two things:

(1) that the signal is speech or other communication signal (because it results from vocal tract resonances, a.k.a. formants) and (2) specific phonetic feature information, particularly manner of articulation (e.g., stop vs fricative vs. nasal, vs. vowel, etc.) and to a certain extent place of articulation (e.g., [b] vs [d] vs [g]). So, phrase scrambling the low-frequency modulation phases effectively screws up the sensory encoding of phonetic information required to trigger the appropriate Deja patterns.

Below is a cartoon illustration of one possible scenario for Deja triggers at the beta oscillation frequency range (associated with segments, like [t]). A variety of modulation patterns emanating from the tonotopic auditory frequency axis is shown on the right across a 4-kHz range. Note the phase (time) disparity in the modulation patterns across frequency. At one level, they reflect the differential distribution of energy across acoustic frequency associated with different phonetic segments. At another level, they reflect the fact that there are three main formants (vocal tract resonances), which are quasi-independent sources. This temporal independence of modulations across acoustic frequency is important, because it informs the auditory system and higher brain centers that this signal was produced by a vocal tract (probably human). On the left, the blue zones indicate where the sensory activity is likely greatest (associated with formants) for consonants, while the green zones are the vocalic concomitants. The red zones show the likely suppression/inhibition following the activation. The time course of the activation + suppression is likely to conform to syllable and theta-rhythm time scales (i.e., ca. 150–250 ms). The full-spectrum modulation pattern shown in upper right looks quite different than the frequency-selective modulations coming out of the narrow-band channels (this figure was created over two years ago, long before Oded's 2011 paper was published – his figure two shows cochlear waveforms (half-wave rectified, but doesn't address the significance of the differential modulation patterns across acoustic frequency).

In the DejaNets framework, at least two (if not more) narrow-band modulation patterns are usually required to trigger the appropriate memory networks. This is why speech usually needs a minimum bandwidth to be intelligible and why narrow spectral slits distributed across the acoustic frequency axis are nearly as intelligible as the entire bandwidth signal, despite 80% of the spectrum being filtered out (Greenberg and Arai, 2004). The crucial property for intelligibility is the sampling of *sufficiently different* modulation patterns. The spectral bandwidth of speech is not nearly as important (it is for robustness, however, where redundancy is crucial in background noise and reverberation).



What is the relationship between the low-frequency modulation patterns of speech and theta oscillations? They are *loosely* correlated, but not precisely. One can think of theta rhythm as one of the basic sampling frequencies of consciousness that reflect data fetching and transmission of information throughout the brain, but which is particularly important for "working memory." The hippocampus appears to be particularly important for short-term memory, and there is a lot of correlated theta activity between that region and the frontal cortex (which is involved in planning, interpretation and decision-making. It is tempting to conclude that theta oscillations are the signature of working memory pure and simple. For reasons described below, this is probably an over-simplification.

David Poeppel and colleagues have recently demonstrated some relationship between theta activity and the ability to process speech (e.g., Luo and Poeppel, 2008). However, not that much is known about the precise relationship. It is also known that individuals who have problems with working memory, both absolute (measured by digit span) and dynamic (as measured by the Trail Making test) have difficulty in understanding speech in noise, regardless of age (Bill Woods of Starkey, personal communication). This finding suggests that there is a vital role in speech comprehension played by memory processes. So, the associated of theta activity with speech perception may be a reflection of memory processes at work during the decoding and comprehension phase of the process.

Another area where theta activity appears to be important is in reading. Good readers have high amounts of theta and delta activity, while poor readers have lower energy in these oscillatory frequency bands

The Importance of Confidence

There appears to be more to theta oscillations than working memory. So far, this other component has been most clearly demonstrated in rats during an odor discrimination task. However, the implications for speech are both clear and profound. Adam Kepecs (Kepecs et al., 2008) has shown that theta activity in the hippocampus is correlated with "confidence" more than with working memory per se. He was able to dissociate the two through a clever operant conditioning paradigm. The magnitude of theta oscillations is more highly correlated with the confidence the rat has about the accuracy of discrimination than with the actual performance level itself. Kiani and Shadlen (2009) have demonstrated the presence of neurons in the monkey parietal cortex that are especially sensitive to their confidence in task performance, suggesting that confidence-sensitive neural responses is potentially a widespread phenomenon.

Why is "confidence" so important for speech? Because without a confidence ranking mechanism, listeners do not possess a way of pruning the large number of alternative interpretations of what is inherently ambiguous sensory input. Information theory focuses on the abstract mathematical computation of alternatives but does not address what is done with this information. In this sense, information theory is incomplete, missing a vital component of the cortical processes involved in processing sensory input and interpreting it in light of what has been previously encountered (i.e., Deja Networks).

When speech begins, ambiguity is high. This is why isolated words, or the beginning of sentences are not decoded nearly as well as the same words embedded in sentential contexts. Add visual cues and some semantic context, and suddenly the task of decoding the acoustic signal becomes much simpler. If the listener's confidence is low, then he/she has a poor path to interpreting what follows, because it is unclear how to interpret this incoming data. For the interpretation process to successfully proceed, confidence must be high. If it's not, then the process has to begin anew (as happens when a listener fails to decode an utterance when originally spoken and requests a repeat – if the speaker delays responding a few seconds, the listener will often decode without further assistance. The lattice of potential interpretations has been "re-scored" and the listener's confidence in the interpretation boosted). The heard of hearing have two, interleaved problems – (1) the sensory triggers are often inadequate and (2) the confidence that the appropriate interpretation has been made is often shattered. This is particularly acute for the elderly who may not possess the cognitive mechanisms required to fill in and compensate for the faulty sensory input.

A comprehensive, detailed description of DejaNets and its implication for speech processing and understanding will be submitted for publication in mid-April, 2012.

Potential Applications for the Project's Research Findings

This research project spanned several fields and methods. It began as a project focused on developing a new form of speech synthesis (STRAIGHT TALK) and concluded with a focus on brain rhythms underlying speech. In between, cross-spectral integration of phonetic information was studied.

What is the underlying them for this broad range of research? That judicious study of brain mechanisms underlying speech and hearing can be used to build useful technology. Below is a brief summary of the potential applications of the project's key components.

(1) STRAIGHT TALK (speech synthesis)

Despite the failure of the STRAIGHT TALK project, a lot was learned about how a special-purpose vocoder, such as STRAIGHT, could potentially be controlled through linguistic input. At present, STRAIGHT lacks a linguistic interface. The only way to generate novel material is through morphing between two different utterances. A better approach would be to map linguistic parameters associated with spoken words into a form "understandable" by STRAIGHT. The project was directed towards this end using phonetic features and the modulation spectrum. The problem was two fold. First, there was no easy mathematical way to combine STRAIGHT with Les Atlas' Complex Modulation Spectrum. We now realize that there is a straightforward (no pun intended) way to do this by truncating the modulation phase to the granularity of the glottal period (ca. 10 ms for male speakers, and 5 ms for female speakers). This insight could make CMS representations within STRAIGHT feasible. I've mentioned this to Hideki Kawahara, STRAIGHT's creator. He seemed both surprised and interested. So perhaps there will be follow-up work in Japan by Hideki and his students.

The other problem confronting STRAIGHT TALK is the mapping of phonetic features to modulation spectral parameters. The project conducted with Thomas Christiansen of the Technical University of Denmark was designed to help overcome this problem by providing an initial set of data linking modulation frequencies to specific phonetic features. Much work remains to be done, but our research demonstrated that there are distinct regions of the modulation spectrum associated with Manner and Place of Articulation as well as Voicing.

The Mach1 system could also help with the development of STRAIGHT TALK, because it shows that selective time compression and expansion of the speech signal can potentially enhance intelligibility. In my view, STRAIGHT is the optimum platform for developing such a system because of its robust way of handling the details of the speech waveform. Done correctly, it is difficult to tell that the signal has been manipulated, so natural does it sound. We return to the potential use of a Mach1-like system below.

(2) Cross-Spectral Integration of Phonetic Information

Much of hearing-aid research is predicated on Articulation Theory and its metric the Articulation Index (AI). The AI assumes that integration of speech information is linear across the acoustic-frequency spectrum, and this assumption forms the basis of much of auditory research focused on developing better hearing aids and cochlear implants. The problem with the AI and its underlying theory is that it is not sufficiently fine-grained. It examines consonant recognition rather than the underlying phonetic constituents. Only by investigating auditory integration at a finer-grained level does the synergistic basis of cross-spectral integration become apparent. Place of articulation information combines across frequency superlinearly. This means that adding additional information about this feature provides far more information gain than would be predicted on the basis of linear integration.

Why is this result important? It provides a way for improving hearing aids through a focus on cross-spectal, super-linear integration of acoustic-phonetic features. Place of articulation provides crucial information for distinguishing among words. Any processing performed that enhanced Place information would therefore result in improved speech intelligibility.

Place of articulation information is highly correlated with visual speech cues, particularly those associated with the movement of the lips, tongue and teeth (Grant, 1998). It is well known that such visual speech cues hugely improves intelligibility, adding as much as 15 dB of signal-to-noise ratio benefit (equivalent to the difference between 10% and 90% intelligibility). The current version of the AI, known as the Speech Intelligibility Index (SII) explicitly excludes visual cues from its framework. This is because the SII does not know how to deal with such super-linear, synergistic information. It suggests that the quantitative foundation of the SII is in radical need of revision.

(3) DejaNets (The role of multi-time-scale oscillations in speech processing and understanding)

There are many potential applications of a memory-based speech theoretical framework, such as DejaNets.

The sensory-decoding framework that has dominated speech and hearing research for the past 90 years (since Harvey Fletcher's pioneering work at Bell Labs) does not provide a comprehensive or deep theoretical foundation for understanding how the brain processes and understands spoken language. Nor does it provide any principled insight as to why spoken language possesses the structure it does, nor account for the wide range of languages around the world (as well as their similarities). Articulation Theory and its ilk are "shallow" theories that only attempt to account for a narrow range of phenomena on a superficial level. They don't try to link all levels of spoken language into a unified theoretical framework. In this sense, the theory associated with speech and hearing is not truly theory in the sense that physicists apply the term.

A drawback of shallow theory is the inability to link disparate phenomena into a coherent perspective. Hence, the applications emanating from such theory is narrow and often not very effective (witness current hearing aid technology, which is of limited utility in noisy backgrounds).

DejaNets is designed to be a "deep" theory, in that it attempts to link many different levels of linguistic behavior and across a broad range of speech phenomena. In this sense, it aspires to be a "theory of nearly everything" in the same way that quantum mechanics (and more recently string theory) try to unify many different levels of the physical world.

Within this deep theoretical framework, many linkages can be made between what otherwise appear as unrelated phenomena.

For example, DejaNets predicts that dyslexics don't have a visual problem per se, but rather a problem *parsing* the speech signal that makes it difficult for them to reliably associated orthographic characters with specific sounds in the context of syllables and words. A corollary of this prediction is that dyslexics, although usually of normal hearing, have significant trouble understanding speech in background noise. Two weeks after this prediction was made, Usha Goswami, a reading specialist at Cambridge University brought such a study (by Jo Ziegler) to my attention, asking what I made of it. She was stunned to learn that I had predicted the result two weeks earlier in my preparation for the aborted STTR proposal. Usha became convinced of the utility of the oscillatory perspective after I sent her a copy of the Ghitza and Greenberg (2009) study. She subsequently wrote a Trends in Cognitive Neuroscience article about how cortical oscillations could account for many aspects of reading difficulties (with acknowledgement to me and David Poeppel for our help in mapping out the connection).

As mentioned elsewhere in this report, the oscillatory perspective encapsulated in DejaNets could potentially help improve hearing-aid technology by selective time compression and expansion of the input signal in order to give the hard of hearing extra time to process and decode the most informative parts of speech. This approach could also be applied to easing the learning of foreign languages and for helping children with language delays to overcome such impediments.

The technology could also help normal-hearing adults understand speech much better in acoustically challenging conditions such as often occur during warfare or other emergency situations. The range of potential applications is broad, because there are many circumstances in which spoken language is difficult to understand. Any technology that enhances intelligibility would be of enormous benefit to verbal communication.